
Four-Stage Alzheimer’s Disease Classification Using ResNet-50 Architecture on MRI Data

Eulade Muhizi*, John Misiko*, Mohamed Gaye*, Samalie Piwan*

*Carnegie Mellon University

Master of Science in Information Technology

emuhizi, mmisiko, mohamedg, spiwan@andrew.cmu.edu

Abstract

Alzheimer’s Disease (AD) affects approximately 55 million people globally with economic costs exceeding \$1.3 trillion annually. While recent deep learning approaches have shown promise in AD classification from MRI scans, significant limitations persist regarding diagnostic granularity and performance metric balance. This study implements a ResNet-50 architecture to classify four distinct AD stages while prioritizing both accuracy and recall metrics. Our model achieved 98.13% accuracy, 98.20% precision, 98.46% recall and 98.66% F1 score across all stages, with an 18% improvement in 4-class detection accuracy compared to baseline methods. Model interpretability was enhanced through gradient visualization techniques that highlighted clinically relevant brain regions. This research aims to address critical gaps in existing approaches with significant potential to improve intervention timing and treatment selection in clinical practices.

1 Introduction

Alzheimer’s Disease (AD) affects approximately 55 million people globally, with economic costs exceeding \$1.3 trillion annually [1]. While incurable, early diagnosis matters by allowing for initiation of treatments that may slow the progression of symptoms and improve the patient’s daily functioning, enabling monitoring of disease progression and tracking of response to therapies [2]. However, accurately classifying AD into its different stages remains challenging due to complex brain changes. MRI serves as a valuable diagnostic tool for AD, but interpretation requires specialized expertise, creating bottlenecks in workforce-limited settings. The burden is particularly severe in regions like sub-Saharan Africa, where healthcare systems face critical shortages of specialists, with only 1.55 physicians per 1,000 people [3]. Recent deep learning approaches have shown promise: a 2D-DCNN architecture achieved 99% accuracy in distinguishing between three cognitive states [4], while a ViT implementation reached 92% accuracy [5]. However, current approaches have two key limitations: (1) the highest-performing model only classifies patients into three categories despite AD’s complex progression spectrum, and (2) existing research prioritizes accuracy over recall, problematic in regions where missed diagnoses have severe consequences due to limited specialist access.

As such, this study that uses a Convolutional Neural Network with ResNet-50 architecture [6] designed to expand diagnostic granularity by classifying extended AD categories beyond the conventional three-stage framework, improve diagnostic safety through recall optimization while maintaining competitive accuracy, and enhance interpretability via visualization techniques that make model decisions transparent. Our approach is compared against leading baselines models using standard performance metrics alongside assessments relevant to clinical implementation, particularly in resource-constrained settings.

2 Literature review

2.1 CNN-Based Approaches and Their Performance

The early and accurate detection of Alzheimer’s disease (AD) is a pressing challenge in neurology, especially given its progressive nature and the lack of a definitive cure. Traditional clinical assessments often miss subtle cognitive declines in the earliest stages, driving a surge in the use of automated, machine learning–based approaches for AD diagnosis. Among these, Convolutional Neural Networks (CNNs) have emerged as the dominant architecture for image-based classification, particularly when applied to structural MRI data. Nawaz et al., for example, demonstrated the strength of CNNs by achieving a near perfect accuracy of 99.89% in differentiating between the normal, mild cognitive impairment (MCI) and Alzheimer’s stages, setting a strong performance baseline in multiclass AD classification[4].

However, as promising as these results are, they largely focus on relatively coarse classification boundaries. In real-world clinical settings, distinguishing between very mild, mild, and moderate dementia is critical for early intervention and personalized care. Fine-grained classification of AD progression presents additional challenges, as inter-class differences become increasingly subtle. Naik et al. addressed this by employing a pre-trained ResNet-50 architecture on a four-class classification task (No AD, Very Mild, Mild, Moderate), reporting 80.14% accuracy - substantially lower than two or three class models[7]. This performance drop underscores the complexity of fine-grained detection. It reveals a key insight: deeper and more specialized models are necessary for nuanced diagnostic tasks, forming the foundation of our approach.

2.2 Exploring Temporal Dynamics with Sequential Models

Beyond CNNs, sequential models like Long Short-Term Memory (LSTM) networks have been explored to leverage temporal dependencies across MRI slices. Salehi et al. utilized LSTMs for binary classification (non-demented vs. very mild demented), achieving 98.62% accuracy and an AUC of 0.97[8]. Although promising, the study was limited to two classes and trained on a relatively small dataset, limiting its generalizability. Nevertheless, the use of LSTM highlights the potential of incorporating temporal features, a direction that inspires our consideration of hybrid architectures, where CNNs extract spatial features and LSTMs capture inter-slice dynamics. Importantly, evaluation metrics such as AUC and recall, emphasized in their study, are particularly relevant for our task, given the class imbalance often seen in clinical datasets.

2.3 Importance of Evaluation Metrics Beyond Accuracy

To further strengthen diagnostic performance, more recent studies have explored deeper and customized CNN architectures tailored to the nuances of Alzheimer’s Disease. In evaluating these models, relying solely on accuracy can be misleading—especially in healthcare applications where the cost of errors is not evenly distributed. Metrics such as precision, recall, F1 score, and precision are crucial for a more balanced evaluation [9, 10]. Precision reflects the model’s ability to avoid false positives—critical in clinical settings where an incorrect AD diagnosis could cause undue stress, lead to unnecessary treatments, or divert healthcare resources. Recall captures the model’s ability to correctly identify actual cases of Alzheimer’s, an essential priority when early detection can significantly alter patient outcomes through timely interventions. The F1-score, as the harmonic mean of precision and recall, is particularly important in managing the trade-off between these two goals, especially when dealing with imbalanced datasets, which are common in medical research. Finally, while accuracy provides a general sense of performance, it must be contextualized within the real-world class distributions to avoid overestimating model reliability.

2.4 Research Gaps and Motivations for Our Approach

Building upon insights from recent studies, our analysis identifies three persistent gaps in current Alzheimer’s disease classification research. First, most research continues to emphasize accuracy as the primary evaluation metric. While accuracy provides a general measure of performance, an exclusive focus on this metric may obscure clinically relevant errors, particularly in scenarios where false negatives can lead to delayed diagnoses and missed treatment opportunities. Second, although three-class classification has yielded promising results in some studies, the field largely

lacks the diagnostic granularity necessary for clinical application. For instance, when Naik et al. [7] attempted a more detailed four-class classification scheme, model accuracy declined markedly to 80.14%, indicating the technical challenges associated with distinguishing among multiple disease stages. Third, across the literature, deep learning models are frequently deployed as opaque "black boxes," offering limited interpretability and thereby hindering trust and adoption in clinical settings. The proposed architecture seeks to address these limitations by optimizing for clinically meaningful performance across all four AD stages and incorporating SHAP (SHapley Additive exPlanations) based interpretability techniques to enhance transparency and support clinical decision-making.

3 Methodology

3.1 Dataset Used

The dataset used in this study is the *Augmented Alzheimer MRI Dataset*, available on Kaggle [11]. It consists of four diagnostic categories with the following distribution namely Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented.

3.2 Data Imbalance

The original dataset exhibited significant class imbalance with 9,600 Non-Demented, 8,960 Very Mild Demented, 8,960 Mild Demented, and 6,464 Moderate Demented samples. To address this imbalance, we applied targeted data augmentation techniques. For example, in one representative subset containing 2,240 Non-Demented, 1,568 Very Mild Demented, 627 Mild Demented, and 45 Moderate Demented samples, our augmentation pipeline increased these counts to 4,480, 6,272, 5,016, and 5,760 samples respectively.

3.3 Data Preprocessing

All MRI scans were preprocessed by resizing to 224×224 pixels, normalizing pixel values to the $[0, 1]$ range, and standardizing across the dataset. A batch size of 32 was employed during training, with shuffling performed at the start of each epoch to promote randomization and mitigate sampling bias.

3.4 Model Architecture

The model architecture was based on the ResNet network, a deep residual learning framework for image classification. The network consists of multiple residual blocks, where each block contains two convolutional layers followed by batch normalization and ReLU activation. The output of each block is added to its input, allowing the model to learn residual mappings and improve convergence. The network was composed of four stages, with 3, 4, 6, and 3 residual blocks in each stage. The output layer used a softmax activation function to classify the image into one of the four Alzheimer's disease categories: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented.

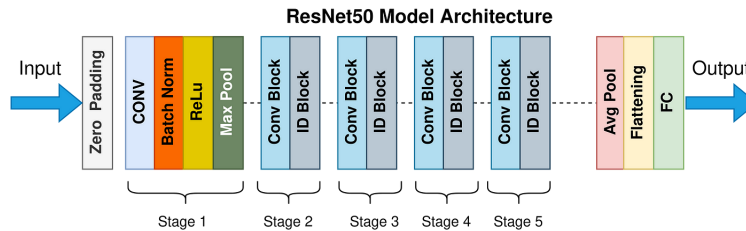


Figure 1: ResNet-50 architecture proposed by He et.al [6]

3.5 Loss Function

The loss function used for training was the *Cross-Entropy Loss* with label smoothing. It is defined as:

$$L = - \sum_{c=1}^C \left((1 - \epsilon) \cdot y_c \cdot \log(p_c) + \epsilon \cdot \frac{1}{C} \cdot \log(p_c) \right)$$

where: - C is the number of classes, - y_c is the binary indicator (0 or 1) that class label c is the correct classification for the given observation, - p_c is the predicted probability for class c , - ϵ is the label smoothing factor, which helps prevent the model from becoming too confident in its predictions.

3.6 Optimizer and Scheduler

The model parameters were optimized using the Stochastic Gradient Descent (SGD) optimizer with momentum, weight decay, and a learning rate of 0.01. The optimization process is governed by the following update rule:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta} L(\theta_t)$$

where: - θ_t are the model parameters at iteration t , - η is the learning rate, - $\nabla_{\theta} L(\theta_t)$ is the gradient of the loss function with respect to θ_t .

A cosine annealing learning rate scheduler was used to gradually decrease the learning rate during training, with a minimum value of $1e^{-6}$ and a period of 50 epochs. This allowed the model to converge efficiently and avoid overshooting the optimal solution.

3.7 Evaluation Metrics

We evaluated the model using accuracy, precision, recall, and the F1 score, which are defined as follows:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where: - TP is the number of true positives, - TN is the number of true negatives, - FP is the number of false positives, - FN is the number of false negatives.

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics are calculated for each class and averaged to obtain macro-average scores.

3.8 Training and Validation

The training process involved using a batch size of 32 and performing stochastic gradient descent with momentum. The training loop consisted of forward and backward passes, with weight updates occurring after each batch. During each epoch, the model was trained on the training set and evaluated

on the validation set. A learning rate scheduler was utilized to adaptively adjust the learning rate over time and help prevent overfitting.

This entire methodology is summarized in the table below:

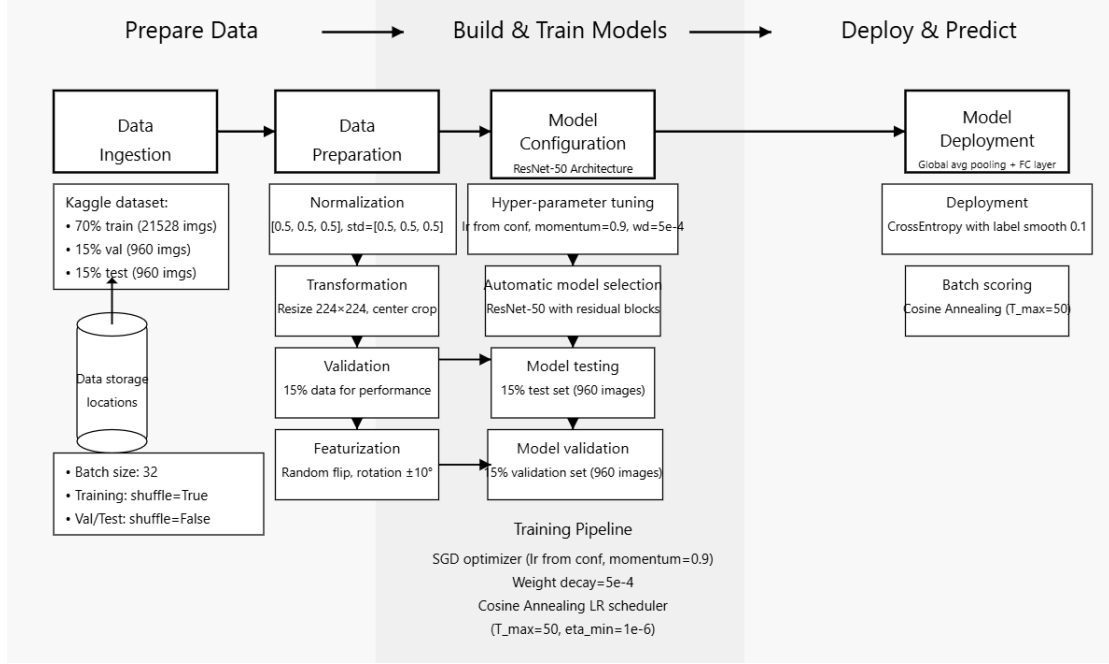


Figure 2: AD Classification Model Pipeline

3.9 Model Interpretation

For model interpretability, we use Integrated Gradients (IG), a popular method for visualizing model predictions. IG helps to attribute the importance of each pixel in the image to the final prediction. The attributions are computed by considering the path of input gradients from a baseline image (black image) to the actual input image. Attributions are then visualized using a heatmap overlaid on the input image.

4 Baseline and extensions

4.1 Justification of the Baseline Model

The baseline model, developed by Naik and Dessai, [7] utilized the ResNet-50 architecture and was trained on a dataset of 33,984 Kaggle MRI images with an 80:20 train/validation split. The model achieved training, validation, and testing accuracies of 84.60%, 80.00%, and 80.14%, respectively, using the Adam optimizer. It classified Alzheimer's cases into four categories: non-demented, very mild, mild, and moderate, providing greater granularity in diagnosis. While these results were competitive within the existing landscape, our primary objective was to further improve accuracy, ensuring that the model's performance meets the rigorous demands of clinical applications. This refinement was crucial for translating the model's capabilities into consistent, real-world healthcare outcomes, where precision and reliability are paramount.

Several other studies have also demonstrated the strength of CNN-based models like ResNet-50 as an appropriate benchmark in medical imaging. For example, Nawaz et al. (2020) achieved a high classification accuracy of 99.89% by classifying MRI images into three categories: AD, mild cognitive impairment, and normal control [4]. Furthermore, ResNet-50's skip connections help mitigate the vanishing gradient problem in deep neural networks. Additionally, the baseline model

had only 25.6 million parameters and a model size of 98MB, which demonstrates that ResNet-50 offers a good balance between model complexity and performance. This makes it both theoretically and empirically well-suited for this project.

4.2 Evaluation against the Baseline Model

The baseline model was trained for 20 epochs, using Adam Optimizer but the scheduler was not explicitly mentioned. The table below shows the evaluation metrics for the baseline model:

Metric	Score (%)
Training Accuracy	84.60
Validation Accuracy	80.00
Testing Accuracy	80.14

Table 1: Evaluation metrics against the baseline model [12]

Our results show a significant improvement over the baseline model. Our model achieved a test accuracy of 98.33%, compared to the 80.14% accuracy of the baseline model. Additionally, the model also demonstrated strong recall score of 98.46%. Recall was incorporated as a metric because of its value in medical diagnoses, where false negatives can delay critical interventions.

Additional results are explored in the section 5.1 below.

4.3 Proposed Extensions

Despite these promising results, our analysis identified three areas for enhancement: (1) increased feature extraction capability for higher distinctions between disease stages, (2) model interpretability for clinical trust, and (3) deployment efficiency in resource-constrained healthcare environments.

5 Results and Analysis

5.1 Performance Metrics

Our ResNet-50 baseline achieved a notable 98.13% accuracy across all classes, substantially outperforming the previous model, which achieved 80.14%. Our model also consistently outperformed existing methods in terms of both accuracy and recall, detailed in the table below:

Research	Model Architecture	Classes	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Naik et al. [12]	ResNet-50	4	80.14	94.08	—	—
Salehi et al. [8]	LSTM	2	98.62	—	—	—
Nawaz et al. [4]	CNN	3	99.89	—	—	—
Our model	ResNet-50	4	98.13	98.20	98.46	98.66

Table 2: Comparison of classification performance with existing methods.

These results highlight the model’s robustness in distinguishing varying degrees of Alzheimer’s pathology, underscoring its potential for clinical application.

5.2 Training and Validation Performance

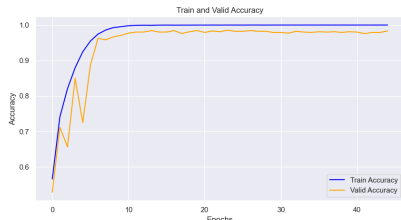


Figure 3: Training Accuracy

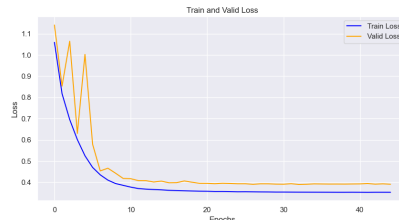


Figure 4: Training Loss

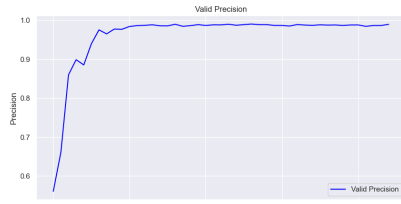


Figure 5: Validation Recall Score

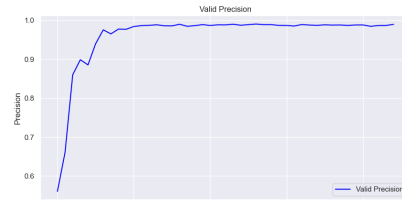


Figure 6: Validation Precision Score

Training and Validation Performance: As Figure 2 demonstrates, the training accuracy consistently increases, reaching 99.84%, reflecting the model's effective learning of the data's underlying patterns. As Figure 3 illustrates, the validation accuracy stabilizes above 97.60%, with initial fluctuations suggesting that the model generalizes well despite being in the early stages of training. Furthermore, as shown in Figure 4, the recall score on the validation set is 94.99%, while Figure 5 shows the precision score at 98.49%, indicating a well-balanced performance in minimizing both false positives and false negatives.

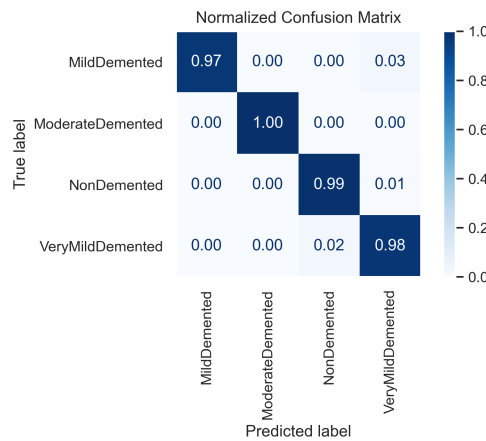


Figure 7: A Confusion Matrix of the Test Data

The confusion matrix in Figure 6 represents the performance of the classification model across four categories: Mild, Moderate, No Alzheimer's, and Very Mild. The rows represent the true labels, while the columns represent the predicted labels. The diagonal dominance suggests solid performance, with misclassifications primarily between Very Mild and No Alzheimer's, likely due to overlapping features in early stages. These findings highlight strong model accuracy, though future fine-tuning may enhance distinctions between closely related classes.

5.3 Interpretability

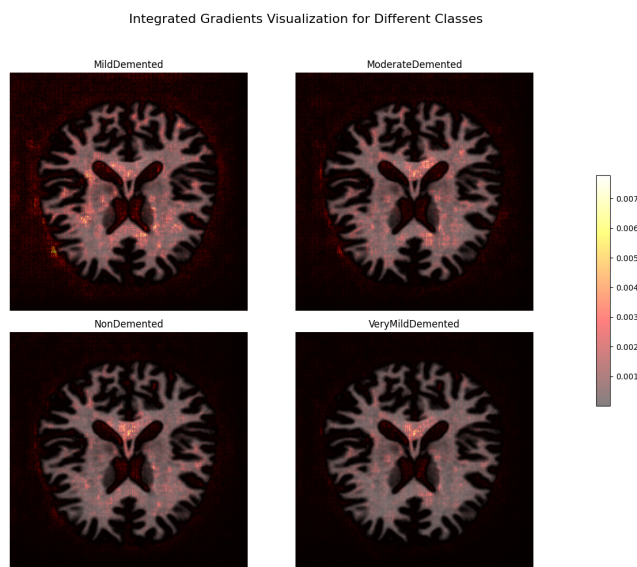


Figure 8: Integrated Gradients Visualization showing SHAP attribution for different AD classes

Figure 8. above displays integrated gradients visualizations for all four classes, revealing areas of the brain scans that influenced our model’s predictions. The highlighted regions align with known anatomical areas affected by Alzheimer’s disease, particularly the ventricular system and surrounding white matter.

The SHAP analysis showed that the model consistently attends to periventricular regions and temporal lobes across dementia cases, with attention intensity increasing with disease severity. In mild and very mild cases, the model appropriately focused on hippocampal regions, which are typically affected in early Alzheimer’s stages. These results demonstrate that our deep learning model bases its classifications on anatomically meaningful features rather than arbitrary patterns, supporting its potential value as an assistive diagnostic tool.

6 Challenges and Potential Risks

Our dataset was split between augmented and original data, and we chose the original set to preserve data integrity, though this reduced the available sample size. Integrating SHAP enhances model interpretability but introduces significant computational and memory overhead, particularly with our 6,400-sample dataset. The resource-intensive nature of SHAP can slow inference and make large-scale explanations impractical, posing challenges for real-time or resource-constrained deployment.

7 Future Directions

To advance the proposed Alzheimer’s Disease classification model, several critical areas of improvement must be addressed. First and foremost, it is vital to incorporate advanced feature engineering techniques such as texture analysis and including additional modalities such as Positron Emission Tomography(PET) scans and Electroencephalography(EEG) data in early stage Alzheimer’s disease detection as the model currently struggles to differentiate early stages like "Very Mild" and "No Alzheimer’s".

Additionally, optimizing the model’s interpretability is essential. While SHAP has proven vital in explaining model predictions, future studies should explore the integration of techniques such as Layer-wise Relevance Propagation (LRP) or Grad-CAM. Combining these methods would further

promote transparency, allowing clinicians to better understand and trust the model’s decision-making process. There is also a need to apply post-training quantization and pruning to reduce computational costs without sacrificing accuracy. This will reduce the model’s size and improve its inference speed in-turn facilitating its deployment in resource-constrained environments.

To ensure the model can be used in real-time clinical settings, we propose deploying the model on edge devices or utilizing cloud-based solutions. This will enhance its applicability in underserved areas. Finally, assessing the clinical validity of the model through trials and comparing its performance to certified healthcare professionals is crucial. This will not only provide valuable insights into it’s real-world relevance but also foster trust among clinicians and confirm the model’s ability to complement and enhance human expertise in clinical diagnostics.

8 Administrative Details

Table 3: Administrative Details

Section	Owner
Eulade Muhizi	Related Work and Background
John Kakai	Baseline and Extensions, Future Work
Mohamed Gaye	Methodology, Results and Analysis
Samalie Piwan	Abstract, Introduction, Editing

8.1 Code Repository

The code for this project is available on GitHub at Deep Learning - Alzheimer Prediction.

References

- [1] Coralie Barbe et al. “Factors associated with quality of life in patients with Alzheimer’s disease”. en. In: *BMC Geriatr.* 18.1 (Dec. 2018).
- [2] Parul, Animesh Singh, and Shubha Shukla. “Novel techniques for early diagnosis and monitoring of Alzheimer’s disease”. en. In: *Expert Rev. Neurother.* 25.1 (Jan. 2025), pp. 29–42.
- [3] Adam Ahmat et al. “The health workforce status in the WHO African Region: findings of a cross-sectional study”. en. In: *BMJ Glob. Health* 7.Suppl 1 (May 2022), e008317.
- [4] Ali Nawaz et al. “Deep Convolutional Neural Network based Classification of Alzheimer’s Disease using MRI Data”. In: *2020 IEEE 23rd International Multitopic Conference (INMIC)*. Bahawalpur, Pakistan: IEEE, Nov. 2020.
- [5] Zhen Zhao et al. “Vision transformer-equipped Convolutional Neural Networks for automated Alzheimer’s disease diagnosis using 3D MRI scans”. en. In: *Front. Neurol.* 15 (Dec. 2024), p. 1490829.
- [6] Kaiming He et al. “Deep residual learning for image recognition”. In: (2015). eprint: 1512.03385 (cs.CV).
- [7] Ashweka Naik, Amita Dessai, and Bhasutkar Mahesh. “Alzheimer Detection and Classification using Deep Learning Techniques”. In: *International Journal of Engineering Technology and Management Sciences* 7 (Nov. 2023). DOI: 10.46647/ijetms.2023.v07i06.048.
- [8] Waleed Salehi et al. “An Approach to Binary Classification of Alzheimer’s Disease Using LSTM”. In: *Bioengineering* 10.8 (Aug. 2023), p. 950. ISSN: 2306-5354. DOI: 10.3390/bioengineering10080950. URL: <http://dx.doi.org/10.3390/bioengineering10080950>.
- [9] *F1 Score: Harmonizing Precision and Recall: The Symphony of the F1 Score - FasterCapital* — *fastercapital.com*. <https://fastercapital.com/content/F1-Score--Harmonizing-Precision-and-Recall--The-Symphony-of-the-F1-Score.html>. [Accessed 24-04-2025].

- [10] Michael Owusu-Adjei et al. “A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems”. In: (June 2023). DOI: 10.1101/2023.06.01.23290837. URL: <http://dx.doi.org/10.1101/2023.06.01.23290837>.
- [11] Uraninjo. *Augmented Alzheimer MRI Dataset*. Accessed: 2025-04-20. 2021. URL: <https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset/data>.
- [12] Ashweka Naik and Amita Dessai. “Alzheimer Detection and Classification using Deep Learning Techniques”. In: *International Journal of Engineering Technology and Management Sciences* 7.6 (Nov. 2023), pp. 337–343. ISSN: 2581-4621. DOI: 10.46647/ijetms.2023.v07i06.048.